Words: 4330

Methods and applications of automatic speech recognition

University of Oulu
Department of Information Processing Science
Bachelor's Thesis
Miko Palojärvi
23.6.2021

Abstract

This thesis is an examination of automatic speech recognition in the form of a narrative literature review. Both past and present methods, and the applications of automatic speech recognition were looked at and examined.

Prior research used for sources in this thesis consists of a wide variety of technical conference papers and journal articles on methods of automatic speech recognition, which has seen a lot of advancements throughout the years, and compilations of knowledge on both methods and applications in the form of books and literature reviews.

For methods of automatic speech recognition, three of the seemingly most significant ones that were examined were dynamic time warping, hidden Markov models, and deep neural networks. The latter one, deep neural networks, seemed to be the most advanced and used one currently.

Applications of automatic speech recognition were looked at with groupings based on their desired communication improvement target, improving either human-human communication or human-machine communication. From the first group, speech-to-speech translation and speech summarization were two popular applications that were examined. From the second group, virtual assistants were examined as an application group of its own, being an encompassing name for a general software agent doing tasks in response to human speech.

The research presented on this thesis has the possibility to serve as a basis of future research on the subject of automatic speech recognition. Suggested avenues for this include a quantitative research analysis on either the performance of different methods, privacy aspects of different applications, or approaching the subject from the point of design science research by documenting construction of an automatic speech recognition application using modern methods.

Keywords
Automatic speech recognition, natural language processing, virtual assistant

Supervisor
Nataliya Shevchuk, Postdoctoral Researcher

Abbreviations

ASR	Automatic Speech Recognition
ATS	Automatic Text Summarization
DNN	Deep neural network
DTW	Dynamic time warping
GMM	Gaussian mixed model
HHC	Human-Human Communication
HMC	Human-Machine Communication
HMM	Hidden Markov Model
NLP	Natural Language Processing
S2S	Speech-to-Speech translation
TTS	Text-To-Speech

Contents

Abstract 2
Abbreviations 3
Contents 4
1. Introduction 5
2. Research methods 6
3. Concepts 7
   3.1 Natural Language Processing 7
   3.2 Automatic Speech Recognition 7
4. Methods 9
   4.1 Dynamic time warping 9
   4.2 Hidden Markov models 9
   4.3 Deep neural networks 10
5. Applications 12
   5.1 Human-Human Communication 12
      5.1.1 Speech-to-speech translation 12
      5.1.2 Speech summarization 12
   5.2 Human-Machine Communication 13
      5.2.1 Virtual assistant 13
6. Discussion 15
7. Conclusion 16
   7.1 Limitations 16
References 18
Appendix A. Research Plan 21

1. Introduction

The purpose of this thesis is to examine the methods and applications of automatic speech recognition (ASR). ASR is used more and more in different applications, both as means to enable a more natural way of human-to-computer interaction, and to enhance and sometimes even enable human-to-human communication. Probably the most notable ASR applications are ones currently in the mainstream of many people's lives, which includes virtual assistants like Apple's Siri, Google Assistant and Amazon Alexa, and speech summarization applications that enable automatic video captioning like Google has available in YouTube.

The thought of humans conversing with machines using speech has historically been considered science fiction. Contrary to this, different methods have been in place since the 1960's to make this a reality. Although rudimentary and error-prone, these early methods and the applications they were used in have paved the way for speech recognition to be a big part in the everyday lives of many people.

Research and development in ASR methods and technologies has vastly improved things throughout the years. There are many notable reasons for this. One noteworthy reason, or rather an explanation, is Moore's Law (Moore, 1965), which continues to be somewhat correct in predicting the increases in computational power throughout the years. Another reason is the rising ability to utilize way more data than before in building models for ASR systems because of advancements in cloud computing, big data, and the ever-growing internet as whole. The third reason is the rise of personal devices like mobile phones, tablets, wearables, IOT-devices, and everything else people use to augment their everyday lives with electronics these days. These devices especially have brought speech to the forefront as an additional interaction method to keyboard-and-mouse (Yu & Deng, 2015).

This thesis aims to contribute on the field of ASR by answering three chosen research questions: Where is ASR used in, what approaches, methods, and technologies are used in ASR applications, and finally what are some potential future uses for ASR? The research methods and limitations are further elaborated on in the next chapter. Answering these questions should provide a slim but comprehensive summary on three important aspects of the technology.

The motivation for this thesis comes from a personal interest in the research area as whole. Although my own experience using ASR applications is very limited, the mechanisms that make them work have always been a question in the back of my mind.

The structure of this thesis is as follows: presenting the research methods that were used to obtain data for the thesis, examining and explaining the relevant concepts and terminologies, examining the most popular ASR methods throughout the years, examining the most popular application groups of ASR, discussing findings, and finally concluding everything together in the last chapter.

2. Research methods

The chosen research method for this thesis is general narrative literature review. A narrative literature review is a type of literature review that aims to summarize scattered knowledge on a topic, to a more easily readable form (Baumeister & Leary, 1997). The primary focus of this review is on providing an overview of the knowledge surrounding the topic of ASR. A secondary focus is also put on the historical development of the methods used in ASR. This means examining the most significant milestones throughout the years, starting roughly from 70's all the way to more recent developments. This thesis will also be following and taking use of some of the guidelines for narrative literature reviews that were created by Baumeister and Leary, (1997).

One clear way to approach this was by defining and then trying to answer research questions. The three research questions that were constructed for this thesis during planning were:

RQ1: Where is automatic speech recognition used in?

RQ2: What approaches, methods, and technologies are used in automatic speech recognition applications?

RQ3: What are some potential future uses for automatic speech recognition?

These research questions were chosen, because they summarize three important questions of a technology, that a reader of a narrative literature review on the subject would have; where the technology is used, how does it work, and how could it be used in the future. By answering these questions, this thesis aims to contribute to the field of ASR by providing a narrative report on these important aspects, while at the same time hopefully providing a list of important previous research so that the reader can further add to their knowledge if they desire.

Most material for research was searched for via Google Scholar, and although some material was in different places, most was ultimately accessed at IEEE Xplore. Also a few were found through Elsevier ScienceDirect. At the beginning, most search terms revolved around the terms that were in the research questions. These terms quickly ballooned to many more, as different concepts came up during research. One excellent source for information early on was the book by Yu and Deng (2015). This provided an excellent avenue to finding new sources that were needed to research relevant concepts further.

3. Concepts

Before diving into the subject matter, the basics and some key concepts need to be explained on the field of natural language processing, and ASR itself. Other concepts that relate further into the methods and applications of ASR are explained in their own chapters.

3.1 Natural Language Processing

Natural language processing is a subfield in computer science, that is focused on exploring how computers can be made to understand natural language used by humans. This is done by examining the intricacies on how humans interact with each other through language, and applying these learned techniques and methods to computers, so that they can understand natural language and use it to perform tasks (Chowdhury, 2003). More broadly, natural language processing can be used in a wide variety of different applications, like automatic machine translation, natural language text processing and summarization, artificial intelligence, and automatic speech recognition to name a few (Chowdhury, 2003). The focus of this thesis is on the last listed topic, automatic speech recognition.

3.2 Automatic Speech Recognition

Automatic speech recognition (ASR), also sometimes referred to as speech recognition or speech-to-text, is a term used to describe processes, technologies and methods that enables better human-computer interaction through translating human speech to a more computer readable format (Yu & Deng, 2015). In addition to improving human-computer interaction, ASR also enables applications that enhance, and sometimes even enable better human-human communication.

In the mainstream ASR and things somewhat closely related to it have been in the limelight many times, thanks to popularization from a couple of different avenues. Most notable ones include Stanley Kubrick's film "2001: A Space Odyssey", which had HAL, a central computer aboard a spaceship that understood and spoke the English language naturally, or KITT, the human-like AI-car companion of David Hasselhoff's character in the tv-series Knight Rider. George Lucas also used human-like robots in his Star Wars films, that like HAL, communicated primarily in natural speech, but also existed like humans alongside them. All these renditions of ASR in popular culture brought it to the knowledge of the general population (B. H. Juang & Rabiner, 2005), despite seeming like far away science fiction at the time.

The main task and purpose of an ASR system is to understand patterns in speech. In a more technical way this can be phrased as an ASR system analyzing audio waveforms, and converting information from it (O'Shaughnessy, 2008). This information can then be used in some benefitting way, usually though the purpose is converting the spoken speech to text. In addition to converting speech to text, it is also possible in some cases to retrieve metadata from the speech, that can be useful. This can be for example the language of the speech, information on the speaker like their gender, age, information on their social and regional origins, and even their health and emotional state (Ververidis & Kotropoulos, 2006; Benzeghiba et al., 2007).

Variations in speech naturally complicate the process of ASR. In addition to variations in the speech itself, secondary factors like background noise complicate things even more. The methods to solve these problems and complications have changed and evolved throughout the years.

4. Methods

The main methods that are going to be looked at here are in order of appearance Dynamic time warping, hidden Markov models and deep neural networks. The first two, DTW and HMM are similar in the way that they address non-stationarity and variability in speech signals, but they fundamentally operate in different manners (B.-H. Juang, 1984). The last method that is inspected is deep neural networks. By its nature neural networks and the related field of artificial intelligence are complicated subjects, but nevertheless the focus here is to try and explain the basics on how deep neural networks are used in ASR.

4.1 Dynamic time warping

In the late 1970s, dynamic time warping (DTW) was suggested to be a solution to combat temporal and spectral variance in speech. As mentioned by O'Shaughnessy (2008), this method required high levels of computing power, and it was not easy to determine how much and what speech templates should be used in the process, which has caused DTW to be replaced by more efficient approaches.

In practice, a DTW procedure stretches speech templates to closely match test and reference samples where in a match could be found, thus matching speech to predetermined words or phrases (O'Shaughnessy, 2008). First, a defined distortion measure is used, which needs to have metrics for relevant differences between sound representations. This is then used to find a reference sequence that has the least amount of dissimilarity, which can be either a word or category. This determines the recognition by minimum distortion (B.-H. Juang, 1984).

4.2 Hidden Markov models

In the 1980s, the next big thing for ASR was to move on from speech templates to statistical modeling. This was done with hidden Markov models (HMM). A Markov model is a concept in probability theory, and it means modelling of a randomly changing system. As for hidden Markov models, the definition is a bit more convoluted. As defined by Rabiner and Juang (1986), "An HMM is a doubly stochastic process with an underlying stochastic process that is not observable (it is hidden), but can only be observed through another set of stochastic processes that produce the sequence of observed symbols." (p. 2). What this means in practice for ASR, is that with HMM speech can be modeled with all its variations, in a consistent and statistical way.

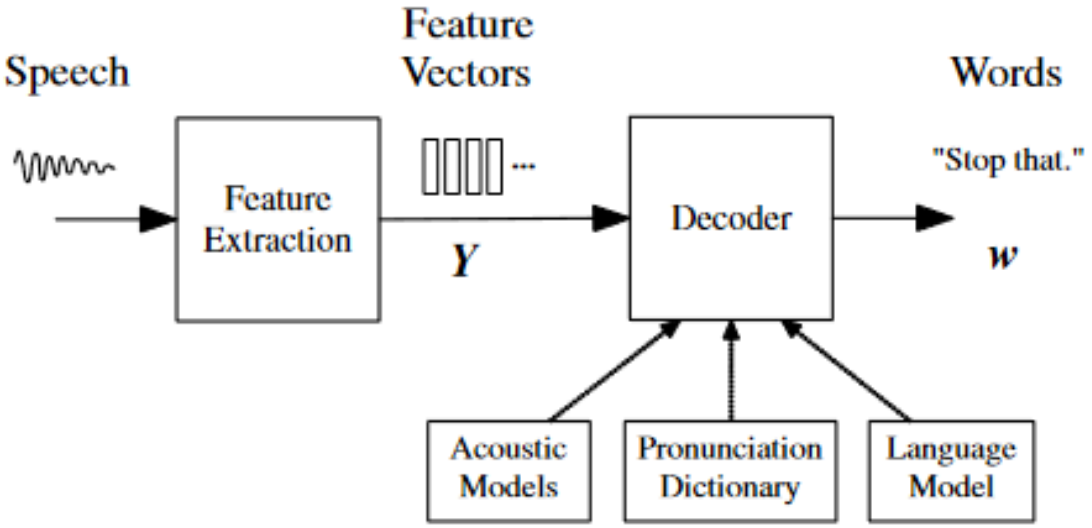

Figure 1. Architecture of an HMM-based speech recognizer system (Gales & Young, 2007).

Figure 1 represents an example model by Gales and Young (2007) of the inner workings of a typical HMM-based speech recognizer system. In this system, first the feature extraction stage converts the input audio waveform into feature vectors. This process aims to transform the speech signal into a more compact representation, while minimizing information loss (Gales & Young, 2007). This data then goes into the decoder, which also gets supporting information from acoustic models, a pronunciation dictionary and language models. The decoder then processes all this information by going through all possibilities of words and removing unlikely words to find out the most likely word sequence, which is the output W in the example model in the figure above (Gales & Young, 2007). Although giving a good general view of an HMM-system, the example is a very simplified version with some assumptions and simplifications made on the used technologies and methods. Gales and Young (2007) also have in their book a chapter called “HMM Structure Refinements”, which provides further descriptions and different ways on expanding on these basic examples.

4.3 Deep neural networks

In a conventional programming scenario, the instructions given to a computer are usually straightforward in the sense that the computer does exactly what it is told to do. These instructions, or the problem that the computer needs to solve, is usually divided into many smaller tasks that the computer has no trouble with. But when the problem is too complicated or too big to solve in a conventional way, a different approach needs to be taken. Neural networks are an answer to this. When using neural networks, the computer does not receive straightforward instructions, but rather it is trained with data, to understand and figure out the solution to the problem.(Nielsen, 2015)

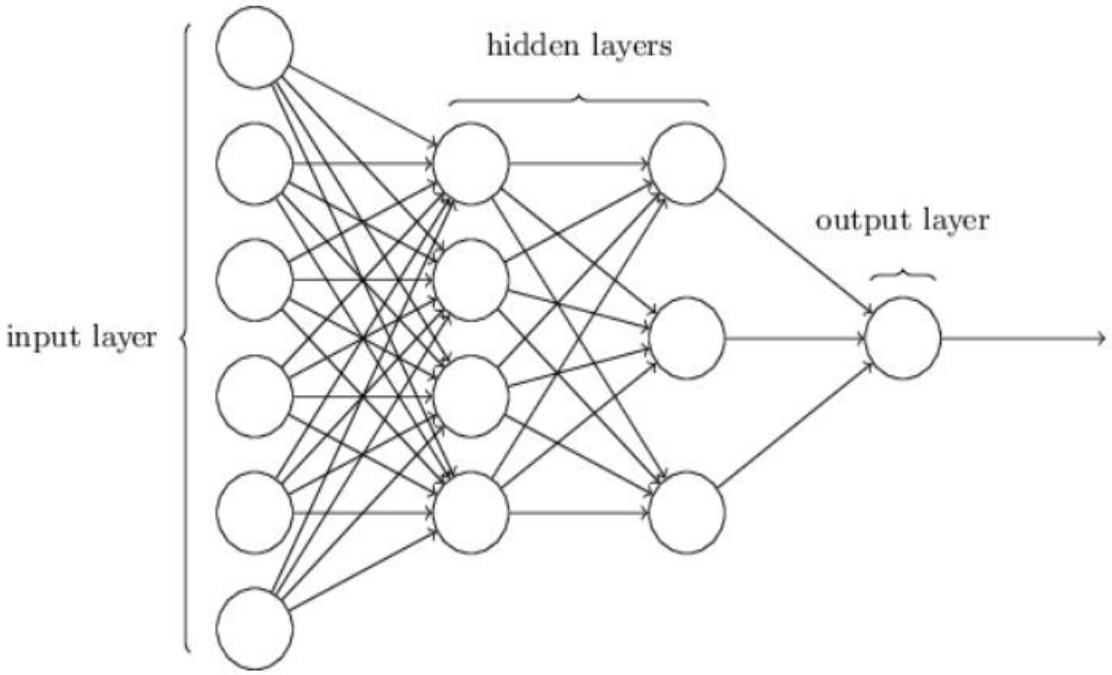

Figure 2. Architecture of a typical neural network (Nielsen, 2015).

Figure 2 represent a typical architecture of a neural network by Nielsen, (2015). The network consists of four layers: the leftmost layer is the input layer, the two middle layers are the hidden layers, and the rightmost layer is the output layer. Each circle in the figure represents a node, or an artificial neuron of sorts. In a simplified way, each node has a predetermined threshold value and a weight associated to it. All these neurons, especially in the hidden layer, combine to a network of passing values through different threshold gates that hold differing weight on the grand scheme of things. This way of achieving output(s) from the passing of input(s) through hidden layers somewhat closely resembles the way a human brain works.

The use of neural networks in speech recognition was a popular subject of interest in the late 80's and early 90's, but at the time it couldn't pass HMM-methods in performance (Deng et al., 2013). Recently though, this has changed. Deng et al., (2013) cite a few reasons for this. The first one is that the neural networks are made to be deeper, which increases potential performance. This is also where the name deep neural network (DNN) comes from. The second reason is the positive effect of sensibly initializing the used weights in the neurons, as discussed for example in Hinton et al., (2012), which brings together the shared views on applying DNN to ASR by 4 different research groups. Next, faster hardware makes training of the neural networks much more effective. And lastly, the performance of DNNs is improved by using more output units that take into account the usage context. (Deng et al., 2013) These days, either a standalone DNN-method, or a hybrid method combining HMM and DNN, seem to be the way to go when developing ASR applications.

5. Applications

In their comprehensive book on ASR, Yu and Deng (2015) describe how ASR applications can be fit into two differing categories: applications that improve Human-Human Communication (HHC), and applications that improve Human-Machine Communication (HMC). Although the terms HHC and HMC do not seem to be fully established in the field of ASR, the grouping and explanation made by Yu and Deng (2015) seems to fit the grouped applications quite well and will be used in this thesis from now on. In addition to the terms not being that well established, both terms (HHC & HMC) are used in similar but overall different concepts, so care needs to be taken in the future when researching this subject area.

5.1 Human-Human Communication

Like mentioned before on this thesis, when researching HHC an obvious pitfall is that the term is used in many different areas of research. Despite this, and although it seeming this term has not been fully established to group ASR-applications, it was chosen for this thesis based on its simplicity, and the good reasoning behind the grouping made overall by Yu and Deng (2015).

5.1.1 Speech-to-speech translation

ASR applications that improve HHC usually do so in a quite straightforward way. The most obvious and best example of this is speech-to-speech (S2S) translation systems.

Figure 3. Components in a typical speech to speech translation system (Yu & Deng, 2015).

S2S translation systems can help people who speak different languages communicate with each other (Yu & Deng, 2015). As shown in Figure 3, the mechanism is basically recognizing speech, and turning it into a computer readable format, translating it with machine translation, then outputting it in the new language with text-to-speech. This is useful for example for travelers in countries where a different language is spoken, who can use S2S to communicate with native people more easily, compared to other methods of translation. This process for S2S-translation generally uses three different technologies: ASR to recognize speech and convert it to text or another computer readable format, machine translation to translate the speech that was converted, and speech synthesis or text-to-speech (TTS) to output the speech in the language that it was translated to (Nakamura, 2009).

5.1.2 Speech summarization

One application that can also be considered to be HHC is speech summarization, or automatic text summarization (ATS). The term used seems to differ depending on the

application method or the context of the application, but the basic idea is the same. As defined by Maybury (1995), “An effective summary distills the most important information from a source (or sources) to produce an abridged version of the original information for a particular user(s) and task(s).” When thought of as an application of ASR, speech summarization could for example mean automatic minute-taking in a meeting, closed captioning, or making abstracts of presentations and speeches (Furui et al., 2001).

Speech summarization is a very important piece of technology in itself, but it can also be applied to be a very useful component in ASR-applications generally. Speech summarization is used especially when the speech that is being processed is spontaneous, meaning the speaker does not read straight from a script or similar ready-made text, but rather the speech itself is naturally unscripted and impromptu. Speech like this makes the spoken words and sentences contains unneeded information like for example filler words, unneeded pauses, unnecessary restarts, interjections, mispronunciations and other things (Ward, 1989). All this makes the speech transcription harder for the ASR-system, so a speech summarization method is needed to extract only important information from speech, while removing unneeded information.

5.2 Human-Machine Communication

Applications grouped into HMC aim to improve, as the name suggest, the communication between humans and machines. One key difference when implementing an HMC application instead of a HHC application is the fact that the speech that is converted into a machine-readable format must also be understood correctly by the machine, for it to be able to use it properly. Yu and Deng (2015) suggest that there are two main components in charge of this; the spoken language understanding-component, which is tasked in finding semantic information on the converted speech, and the dialog manager-component, which handles everything needed for communicating with applications that make use of the speech, in addition to communicating with the other components of the ASR-system.

Yu and Deng (2015) list five categories of applications classified as improving HMC: voice search, personal digital assistant, gaming, living room interaction systems, and in-vehicle infotainment systems. For this thesis however, the specific application group that is being looked at is the virtual assistant. This is because all the categories that were mentioned before have the same basic inner functionalities when examining them from the perspective of ASR, with differences mainly occurring due to the changes in the environment they are used in.

5.2.1 Virtual assistant

Virtual assistant, voice assistant, or intelligent virtual assistant describe basically the same thing: a software agent that can interpret human speech, formulate a response or do a task, and in some cases even respond with TTS (Hoy, 2018). These tasks include for example transcribing and sending an email or a text message, playing music or video, controlling home automation like temperature controls or lights, doing tasks related to navigation, or even ordering products online. A virtual assistant can also respond to questions like “What is the time in Hong Kong?” or “Is it going to rain today”. To answer questions like these, or to be able to achieve tasks set out for it like the previously mentioned ones, the

virtual assistant makes use of many different applications that have been programmed to be compatible with commands that come from human speech.

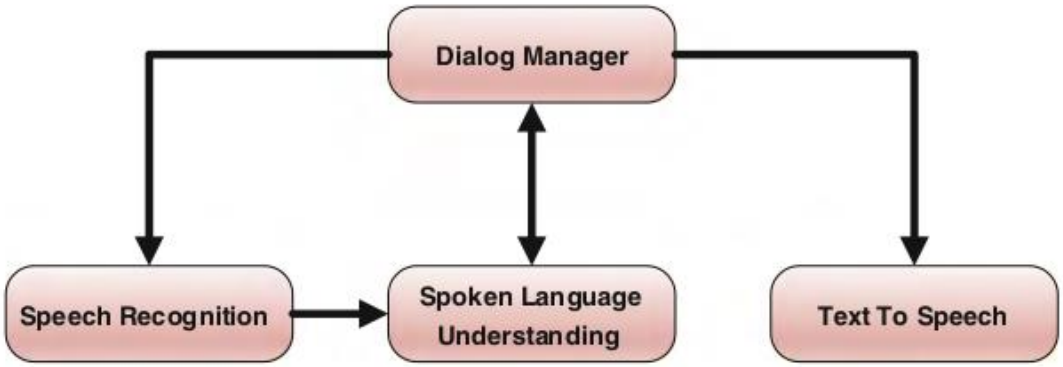

Figure 4. Components in a typical spoken word language system (Yu & Deng, 2015).

Architecturally, virtual assistants usually contain many different components in addition to the ASR-part, in order for it to achieve the functionality of a human-like assistant. Figure 4 illustrates a simplified version of a typical spoken word language system by Yu and Deng (2015), which contains many of the same main components, so it is used here to illustrate a typical architecture of a virtual assistant. The first block is speech recognition, which converts speech to text. The next block, spoken language understanding, is responsible for finding semantic information in the text. After that, the dialog manager conveys this information to other components in the system. The last block, TTS, can be used to convey information back to the user, in a human-like fashion.

6. Discussion

In the beginning of the thesis, three research questions were outlined. The first question was about the applications of ASR, more specifically examining where ASR is used these days. Three major application types were found, S2S-translation, speech summarization and virtual assistant. The first two aim to improve HHC, and while their use-cases may seem niche and uncommon, they seem to be notable applications in the field of ASR. The last application type, virtual assistant, aims to improve HMC, and seems to be the most prominent and visible use of ASR in the world today. In the end, not too much of deep research was done in this thesis on the space of virtual assistants, as the topic goes more into specific companies and their methods of applying ASR.

The second research question relates to this, seeking to learn exactly what are the approaches, methods, and technologies used in ASR. The methods have changed a lot throughout the years, with the three most prominent ones looked at in this thesis being DTW, HMM and DNN. The use of DNN, or neural networks in general seem to have had the biggest impact in increasing performance in ASR applications, and their popularization.

The third research question was about possible future uses for ASR. An answer to this is not nearly as straightforward as the other research question, but possible avenues for the future of ASR, and prominent future research in the subject area is delved into more in the next chapter, conclusion.

7. Conclusion

In conclusion, the three most significant methods of ASR are in DTW, HMM and DNN. DTW seems to be a thing of the past, with only a few closed systems designed for specific tasks using it. Most systems these days use some type of DNN, either it be a system that uses a combination of DNN-HMM, or DNN in addition to some other similar techniques.

The applications of ASR can be grouped differently, and the grouping here was applications that improve HHC, and applications that improve HMC. HHC includes applications like S2S-translation, which can be of great help in improving multilingual dialogue. Another HHC application is speech summarization, which can be used in many different contexts, the most prevalent ones being automatic minute taking, closed captioning, and automatic transcription of presentations and speeches.

HMC applications were grouped into one specific but these days very popular thing, virtual assistants. Virtual assistants enable users to use their voice as an input to different devices, to achieve everyday tasks more easily. Virtual assistants can answer questions and complete a wide range of different tasks given to it.

Overall, it seems that the trends for advancements in ASR is and has been deep learning, deep neural networks, and usage of artificial intelligence. Current applications of ASR all seem to be possible due to these aforementioned methods and technologies, and unless a big shift in different direction happens, the future of ASR is in deep neural networks, and other methodologies that are related to artificial intelligence.

In an article on advancements in deep learning and artificial intelligence by Markoff (2012) in The New York Times, Geoffrey E. Hinton, a leading researcher in the field of deep learning is quoted as saying: “We decided early on not to make money out of this, but just to sort of spread it to infect everybody”. This quote is also said in relation to a rather negative trend of disputes and controversies in intellectual property rights in state of the art-technology fields (Markoff, 2012). Based on this quote, and the rising trend of open-sourcing new ideas and technologies in the field of computer science, one could ascertain promises for a better future in further advancements that should reap benefits for everyone.

Future research on this subject area could for example be a qualitative research analysis on effectiveness of different modern methods of ASR, comparing different performance and quality markers that actually matter in real-world situations. Another possible avenue of future research is documenting the process of designing, developing, and testing of an ASR system with modern methods, possibly using open-source resources. The documentation on the efforts that were made on different parts of the process could provide a valuable resource for even further research on ASR development.

7.1 Limitations

Two possible limitations were thought of during planning for this thesis. One was in relation to difficulties in finding information about approaches, methods, and technologies that specific ASR applications used. This limitation inevitably did not matter that much, as the topic of the thesis strayed more into the realm of examining the methods and applications in general, rather than pinpointing specific methods into specific applications.

The second limitation was dealing with the complexity of the topic. In some ways, this limitation did occur, especially on the ASR-methods part of the thesis. I personally feel that I developed a good general understanding of all the topics that were researched, and a deeper understanding on a few of them. Despite this, some things might not translate to a reader that has little to none of previous knowledge on the topics. Especially the explanation on ASR-methods may not be easily understood without previous experience or familiarization with DTW, HMM or DNN, or things relating to them.

References

Baumeister, R. F., & Leary, M. R. (1997). Writing Narrative Literature Reviews. *Review of General Psychology*, *1*(3), 311–320. https://doi.org/10.1037/1089-2680.1.3.311

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, *49*(10–11), 763–786. https://doi.org/10.1016/j.specom.2007.02.006

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. https://doi.org/10.1002/aris.1440370103

Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8599–8603. https://doi.org/10.1109/ICASSP.2013.6639344

Furui, S., Iwano, K., Hori, C., Shinozaki, T., Saito, Y., & Tamura, S. (2001). Ubiquitous speech processing. *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, *1*, 13–16 vol.1. https://doi.org/10.1109/ICASSP.2001.940755

Gales, M., & Young, S. (2007). The Application of Hidden Markov Models in Speech Recognition. *Foundations and Trends® in Signal Processing*, *1*(3), 195–304. https://doi.org/10.1561/2000000004

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of

Four Research Groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

Hoy, M. B. (2018). Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical Reference Services Quarterly*, *37*(1), 81–88. https://doi.org/10.1080/02763869.2018.1404391

Juang, B. H., & Rabiner, L. R. (2005). *Automatic Speech Recognition – A Brief History of the Technology Development*. 24.

Juang, B.-H. (1984). On the hidden Markov model and dynamic time warping for speech recognition—A unified view. *AT T Bell Laboratories Technical Journal*, *63*(7), 1213–1243. https://doi.org/10.1002/j.1538-7305.1984.tb00034.x

Markoff, J. (2012, November 24). Scientists See Promise in Deep-Learning Programs. *The New York Times*. https://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html

Maybury, M. T. (1995). Generating summaries from event data. *Information Processing & Management*, *31*(5), 735–751. https://doi.org/10.1016/0306-4573(95)00025-C

Moore, G. E. (1965). *Cramming more components onto integrated circuits*. *38*(8), 6.

Nakamura, S. (2009). *Overcoming the Language Barrier with Speech Translation Technology*. 14.

Nielsen, M. A. (2015). *Neural Networks and Deep Learning*. http://neuralnetworksanddeeplearning.com

O'Shaughnessy, D. (2008). Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, *41*(10), 2965–2979. https://doi.org/10.1016/j.patcog.2008.05.008

Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, *3*(1), 4–16. https://doi.org/10.1109/MASSP.1986.1165342

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, *48*(9), 1162–1181. https://doi.org/10.1016/j.specom.2006.04.003

Ward, W. (1989). Understanding spontaneous speech. *Proceedings of the Workshop on Speech and Natural Language - HLT '89*, 137–141. https://doi.org/10.3115/100964.100975

Yu, D., & Deng, L. (2015). *Automatic Speech Recognition*. Springer London. https://doi.org/10.1007/978-1-4471-5779-3

Appendix A. Research Plan

1. Introduction

Automatic Speech Recognition, or ASR, is used to describe technologies and methods that enables better human-computer interaction through translating human speech to a more computer readable format (Yu & Deng, 2015). There are many uses for this, most notably for example virtual assistants like Apple's Siri, Google Assistant and Amazon Alexa.

ASR as a field has developed rapidly throughout the years. In the early days it was considered science fiction for a computer to understand human voice. Further advancements were also hampered by performance limiting factors like insufficient computing power. But advancements in the field of computer science in the 21st century has made huge improvements possible on different ASR approaches (Yu & Deng, 2015; Ghai & Singh, 2012).

2. Research questions and methods

The three main research questions of this thesis are:

RQ1: Where is automatic speech recognition used in?

RQ2: What approaches, methods, and technologies are used in ASR applications?

RQ3: What are some potential future uses for ASR?

Answering these research questions, primarily RQ1 and RQ2, will hopefully generate a slim but a comprehensive thesis on the topic of applications of automatic speech recognition.

3. Limitations

One limitation I foresee is that information may not be available on the approaches, methods and technologies used on some ASR applications. Although through some early research I found some indication that for example Apple is highly open about the methods and technologies that they use.

Another possible problem is the complexity of the topic. Understanding the research material is crucial to this thesis. I probably will not be able to construct a neural network model for my own ASR after this, but a general understanding of the underlying mechanics of different methods and technologies should form during this thesis.

4. Preliminary earlier research

Most of my research thus far has consisted of searching for earlier research material that is related to ASR and its field of study overall. That means that I have not deep dived into the topics yet, as I have tried to construct a general view in my mind of the research area, while constructing appropriate research questions that would be suitable for this thesis.

5. Sources

Ghai, W., & Singh, N. (2012). Literature Review on Automatic Speech Recognition. *International Journal of Computer Applications*, *41*(8), 42–50. https://doi.org/10.5120/5565-7646

Yu, D., & Deng, L. (2015). *Automatic Speech Recognition*. Springer London. https://doi.org/10.1007/978-1-4471-5779-3

6. Timetable

wk. 18: Initial planning, finding thesis supervisor

wk. 19-25: Researching and writing

wk. 26: Finalized version

7. Preliminary table-of-contents

Abstract

Abbreviations

Contents

1. Introduction
2. Concepts
   2.1 Natural Language Processing
   2.2 Automatic Speech Recognition
3. Research methods
4. Findings
5. Discussion
   5.1 Current uses
   5.2 Potential future uses
6. Conclusion

References

Appendix A – Research Plan